ВІДГУК

ОФІЦІЙНОГО ОПОНЕНТА

кандидата технічних наук, доцента Чалої Лариси Ернестівни на дисертаційну роботу Аджіт Пратап Сінгх Гаутама «Інформаційна технологія екстракції бізнес знань з текстового контенту інтегрованої корпоративної системи», що подана на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології

Актуальність теми дослідження. Останнім часом набуває стрімкий розвиток застосування інформаційно-комунікаційних технологій в галузях, що потребують обробки неструктурованої інформації. При цьому виникає необхідність суттєвого вдосконалення методів автоматизованої обробки та зберігання інформації, яка об'єднує факти і знання із різноманітних галузей людської діяльності для вирішення повсякденних завдань. В цьому контексті важливим напрямом теоретичних досліджень і прикладних розробок є створення автоматизованих інформаційних систем обробки даних, що забезпечували б базис для вирішення різноманітних завдань управління. До таких завдань, зокрема, належить створення інформаційного простору підприємств, до якого входять не лише явним чином визначені знання та структуровані дані, що представлені у традиційних базах даних, але й неструктурована інформація, яка може бути присутньою у різноманітних текстових документах та є критичною для прийняття важливих бізнес рішень. У зв'язку з цим, крім методів, що традиційно використовуються для підтримки прийняття бізнес рішень (наприклад, методів аналізу фінансових рядів, що базуються на звітній інформації поточного фінансового й економічного стану компаній), потужним додатковим засобом прийняття бізнес рішень стають знання, які вилучаються з текстів відповідної проблематики. Слід відзначити, що в сучасних умовах переважна більшість технологій обробки текстової інформації базується на традиційних статистико-імовірнісних підходах, що майже не використовують опрацювання смислу. Внаслідок цього системи, які автоматизують процес екстракції знань з великих обсягів текстових даних, що представлені у інтегрованих корпоративних системах (ІКС), мають досить низьку повноту й точність видобування знань з текстового контенту.

Це обумовлює актуальність дисертаційної роботі Аджіт Пратап Сінгх Гаутама, присвяченої питанням розробки та практичного застосування

моделей, методів та інформаційної технології екстракції бізнес знань з текстового контенту інтегрованої корпоративної системи.

Основні представлені в роботі результати були отримані здобувачем на кафедрі інтелектуальних комп'ютерних систем НТУ «ХПІ» при виконанні держбюджетних тем МОН України: «Розробка математичних моделей та методів розв'язання задач інтелектуальної обробки інформації» (ДР № 0108U003926), «Розробка моделей та методів для інформаційно-пошукових, лексикографічних інтелектуальних систем» (ДР № 0111U002258).

Ступінь обґрунтованості та достовірності наукових положень, висновків і рекомендацій, сформульованих в дисертаційній роботі. Обґрунтованість та достовірність наукових положень, висновків і рекомендацій, які сформульовані в дисертаційній роботі Аджіт Пратап Сінгх Гаутама**,** базується на ретельному аналізі науково-технічних джерел за тематикою роботи, коректному визначенні мети й постановці задач дослідження, використанні сучасних методів дослідження, дослідженні та критичному аналізі отриманих результатів, якісному формулюванні отриманих висновків. Теоретичні дослідження, пов'язані зі створенням інформаційно-логічних моделей смислової обробки і методів екстракції бізнес знань з текстового контенту, а також для формування інформаційного простору інтегрованої корпоративної системи, виконано з використанням сучасного математичного апарату (теорії інтелекту, алгебри скінченних предикатів, методу компараторної ідентифікації, методів багатокритеріальної класифікації, методів текстових колекцій та математичної теорії вибірки). Отримані результати не суперечать відомим поняттям і визначенням, а доповнюють і розвивають їх, що підтверджує обґрунтованість наукових положень, висновків і рекомендацій, сформульованих в дисертаційній роботі результатів дослідження.

Достовірність результатів дисертаційної роботи, одержаних здобувачем, підтверджено актами їх практичного використання для розробки підсистем формування фактографічних баз даних й фактографічного пошуку наукових бібліотек НТУ «ХПІ» і Харківського національного університету радіоелектроніки, а також впровадження до навчального процесу на кафедрі інтелектуальних комп'ютерних систем НТУ «ХПІ».

Оцінка змісту дисертації та автореферату. Дисертація складається зі вступу, чотирьох розділів, висновків, списку використаних джерел і додатків.

У вступі обґрунтовано актуальність роботи, сформульовано мету та задачі дослідження, наведено положення, що виносяться на захист, відзначено їх новизну та практичне значення, надано відомості щодо публікацій та апробації результатів роботи.

У **першому розділі** на основі аналітичного огляду інформаційних джерел проаналізовано існуючі задачі і проблеми в області автоматизації корпоративних систем. Підкреслено необхідність поглибленого аналізу даних та підвищення ефективності систем інформаційної підтримки прийняття рішень, а також електронному документообігу і діловодства в ІКС управління територіально розподіленою корпорацією. Вдосконалення таких систем має поєднати стратегію управління підприємством та провідні інформаційні технології, засновані на єдиній програмно-апаратній платформі й спільній базі знань. Відзначено, що до актуальних функцій сучасних ІКС слід віднести накопичування структурованих формалізованих знань, які дозволяють доступно і багаторазово вирішувати реальні виробничі та організаційні завдання на рівні всієї корпорації. Визначені здобувачем у ході аналізу недоліки існуючих автоматизованих систем та можливі шляхи їх усунення дозволили окреслити сукупність задач, які розв’язуються у дисертаційній роботі.

У другому розділі дисертації обґрунтовано вибір математичних засобів для формалізації процесу витягу знань з текстового контенту ІКС. У якості математичного апарату опису дискретних, детермінованих і скінченних об’єктів системи представлення знань, добутих з текстового контенту ІКС, запропоновано використання алгебри скінченних предикатів (АСП). Запропоновано схему компараторної ідентифікації фактів для бізнес аналізу роботи корпорації.

У **третьому розділі** представлено процедури формування інформаційного простору ІКС, що включає відбір концептів, які визначають базові сутності знань корпорації; класифікацію базових сутностей через продукування класів еквівалентності концептів; формування атрибутів сутностей; визначення відношень між сутностями у вигляді *Is-A, Part Of* , *Include* та ін.

Динамічна екстракція знань з текстів потребує ідентифікації сутностей, які включаються у інформаційне поле бізнес інтересів корпорації. Для витягу з тексту імен сутностей, які позначають деякі об’єкти або суб’єкти

дискретного світу, і віднесення їх до певного семантичного класу, що визначає тип його поведінки у триплеті факту, удосконалено метод виявлення актуальної множини класифікованих сутностей предметної області. Для зменшення типів класу використовується кодування, що дозволяє об'єднувати характеристики різних токенів на основі аналізу ступеня близькості оформлення до їх графемного виду. Запропоновано логіко-лінгвістичну модель генерації фактів з текстових потоків ІКС, яка базується на використанні поверхневих граматичних характеристик ідентифікації сутностей дій та атрибутів.

Четвертий розділ присвячено удосконаленню інформаційної технології формування єдиного інформаційного простору бізнес діяльності корпорації, яка включає наступні функції:

- витяг даних, добування інформації і знань з контенту Веб-сторінки;
- пошук і витяг елементів знань, явним чином присутніх у тексті у вигляді твердження або факту, які здійснюються на базі логіко-лінгвістичної моделі генерації фактів з текстових потоків ІКС;
- породження складного знання шляхом узагальнення, яке здійснюється за рахунок структурування відношень фактів бізнес знань ІКС.

Для оцінки ефективності процесу екстракції знань з текстового контенту використовується метод текстових колекцій, який полягає у порівнянні результатів роботи імплементованої технології з деяким еталонним результатом для кожного конкретного набору текстів. Визначення обсягу експериментально досліджуваних текстів проводиться методом математичної теорії вибірки, який визначає механізми формування репрезентативної поворотної вибірки.

Висновки до розділів та за результатами роботи сформульовані достатньо чітко і виразно та відповідають змісту дисертаційної роботи.

Список використаних джерел зі 137 найменувань є досить повним та охоплює сучасні вітчизняні та зарубіжні публікації за темою дисертації.

Зміст автореферату відображає основний зміст дисертації та достатньо повно розкриває наукові положення та практичну цінність роботи.

Наукова новизна основних положень дисертаційної роботи. Наукова новизна основних положень дисертаційної роботи полягає в розробці та дослідженні інформаційно-логічних моделей і методів смислової обробки текстового контенту, що дозволяють підвищити ефективність інформаційної технології екстракції бізнес знань інтегрованої корпоративної системи.

Можна погодитись з висновками здобувача, що наукову новизну мають такі положення дисертації:

- уперше розроблено логіко-лінгвістичну модель генерації фактів з текстових потоків інформаційної корпоративної системи, яка базується на використанні поверхневих граматичних характеристик сутностей, предикатів та атрибутів, що дозволяє ефективно екстрагувати з текстового контенту профільні знання про суб'єкти моніторингу;

- уперше розроблено метод створення інформаційного простору фактів інтегрованої корпоративної системи, заснованої на побудові ієрархічної структури кластерів фактів, яка базується на гіпонімічних відношеннях більш високого порядку узагальнення, що дозволяє структурувати вилучені знання про економічну діяльність корпорації за видами продукції, галузями виробництва, географічному включенню тощо;

- отримав подальший розвиток метод компараторної ідентифікації, який використано для структурування відношень фактів бізнес знань ІКС, реалізація якого дозволяє класифікувати атрибути сутностей за класами відношень за рахунок смислової тотожності триплетів фактів, що об'єктивно визначено компаратором;

- удосконалено метод виявлення актуальної множини класифікованих сутностей предметної області, який відрізняється комплексним використанням лінгвістичних, статистичних й смислових характеристик в наївному байєсівському класифікаторі, що дозволяє класифікувати сутності, які екстрагуються з тексту, за апріорно виділеними типами;

- удосконалено інформаційну технологію формування єдиного інформаційного простору бізнес діяльності корпорації, яка дозволяє за рахунок використання алгебро-логічних перетворень здійснювати породження складного знання шляхом експліцитного узагальнення інформації, що прихована у сукупності часткових фактів.

Значимість отриманих результатів для практичного використання. Практичне значення отриманих результатів полягає у розробці інформаційної технології формування єдиного інформаційного простору бізнес діяльності корпорації. Розроблена технологія включає логіко-лінгвістичну модель генерації фактів з текстових потоків ІКС, метод структурування відношень фактів бізнес знань корпорації, метод виявлення актуальної множини класифікованих сутностей предметної області, а також спеціалізовані етапи Web Content Mining лінгвістичного процесора. Запропоновані у дослідженні математичні моделі можуть бути використані у системах автоматичного

опрацювання текстів, системах вилучення знань, добування інформації (Information Extraction) і розпізнавання сутностей (Named Entity Recognition).

Результати дослідження імплементовані у веб-додаток, який здійснює моніторинг фактографічної інформації заздалегідь визначених компаній та корпорацій. Практичне значення результатів роботи підтверджується їх впровадженням у підсистемах формування фактографічних баз даних й фактографічного пошуку наукових бібліотек НТУ «ХПІ» і Харківського національного університету радіоелектроніки. Використання розроблених у роботі моделей і методів дозволило підвищити ефективність технологій екстракції бізнес знань з текстового контенту за рахунок підвищення середніх значень коефіцієнтів повноти й точності видачі фактографічної інформації.

Теоретичні аспекти дисертаційної роботи використовуються в навчальному процесі на кафедрі інтелектуальних комп'ютерних систем НТУ «ХПІ» при викладанні спеціальних дисциплін «Інформаційно-ресурсне забезпечення лінгвістичної діяльності», «Штучний інтелект: лінгвістичні проблеми», «Автоматизована обробка природної мови» для студентів спеціальності «Прикладна лінгвістика» та при виконанні курсових й дипломних робіт.

Повнота викладення в опублікованих працях основних результатів наукових досліджень дисертації. За темою дисертаційної роботи опубліковано 14 наукових праць, зокрема: 4 статті у наукових фахових виданнях України з технічних наук, 6 статей у наукових періодичних іноземних виданнях, 4 тези доповідей у матеріалах науково-технічних конференцій. В опублікованих працях повністю відображено основні результати дисертаційного дослідження.

Автореферат ідентичний за змістом з основними положеннями дисертаційної роботи, достатньо повно відображає основні її наукові результати, отримані здобувачем, написаний достатньо грамотно та з використанням сучасної наукової термінології. Оформлення дисертаційної роботи й автореферату відповідає діючим вимогам.

Відповідність дисертації паспорту спеціальності. Дисертаційну роботу виконано у відповідності до пунктів 3 (розроблення моделей і методів автоматизації виконання функцій і завдань виробничої й організаційного управління у звичайних і багаторівневих структурах на основі створення і використання нових інформаційних технологій) та 13 (створення

інформаційних технологій для розроблення моделей, методів та інструментальних засобів автоматизації інформаційно-пошукових і телекомунікаційних систем, мереж і засобів інформаційного забезпечення бібліотек, музеїв і архівів (електронні каталоги, автоматизовані робочі місця, комп'ютерна бібліографія, системи автоматизованого імпорту документів тощо)), зазначених в паспорті спеціальності 05.13.06 – інформаційні технології.

Зауваження до дисертаційної роботи.

1. З тексту дисертації не зрозуміло, чи існують інформаційні системи автоматизованої екстракції бізнес знань з текстового контенту, крім розробленої автором; якщо подібні системи існують, то бажано було б дати їх порівняльну характеристику, якщо ж таких немає, то доцільно було б це зазначити.

2. Не цілком зрозуміло, яким чином використовується в інформаційній технології екстракції бізнес знань з текстового контенту формальна модель наївного байєсовського класифікатора типів сутностей предметної області, що запропонована в підрозділі 2.5.

3. Розроблена у третьому розділі дисертаційної роботи концептуальна модель генерації фактів з текстових потоків ІКС потребує більш детального пояснення та обґрунтування, зокрема, слід було приділити більше уваги опису моделі сутностей, предикатів та атрибутів.

4. Запропонований в дисертаційній роботі метод створення інформаційного простору фактів ІКС базується на побудові ієрархічної структури кластерів фактів з урахуванням гіпонімічних відношень високого порядку узагальнення, що є, безумовно, прогресивним напрямком розвитку сучасних інформаційних технологій. В той же час, цей метод не є логічно пов'язаним з питаннями розробки предметних онтологій, формалізації відповідних процедур та їх застосуванням у розробленій здобувачем інформаційній технології.

5. В тексті дисертаційної роботи використовується багато специфічних термінів та понять з галузі лінгвістичних наук, що ускладнює сприйняття роботи та не дає можливості сконцентрувати увагу на предметі дослідження; було б більш доцільно використовувати терміни, які більш поширені у галузі інформаційних технологій, а також більш чітко висвітлити результати власних досліджень.

6. Результати експериментальних досліджень, що наведено в четвертому розділі, містять лише значення показників коректності екстракції бізнес знань з текстового контенту за запропонованими методами. Цікаво було б також оцінити час виконання відповідних обчислень та обсяг пам'яті, що потребується. Це б дозволило обґрунтувати можливість використання запропонованої технології для оперативної обробки значних обсягів даних в реальному часі.

Однак ці зауваження суттєво не впливають на загальну позитивну характеристику дисертації, що має визначені вище актуальність, наукову новизну і практичну значущість.

ВИСНОВОК

Дисертаційна робота Аджіт Пратап Сінгх Гаутама «Інформаційна технологія екстракції бізнес знань з текстового контенту інтегрованої корпоративної системи», за своїм змістом відповідає паспорту спеціальності 05.13.06 – інформаційні технології. Дисертаційна робота є завершеним науковим дослідженням, спрямованим на розробку формальних моделей та інформаційної технології екстракції бізнес знань з текстового контенту інтегрованої корпоративної системи. Вважаю, що за актуальністю обраної теми, достовірністю і обґрунтованістю висновків, новизною досліджень, значимістю отриманих результатів для науки і практики дисертаційна робота повністю відповідає п. 11, а автореферат п.13 «Порядку присудження наукових ступенів», що затверджено постановою Кабінету Міністрів України № 567 від 24.07.2013 р., щодо кандидатських дисертацій, а здобувач, Аджіт Пратап Сінгх Гаутам, заслуговує присудження наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології.

Офіційний опонент:
кандидат технічних наук, доцент,
доцент кафедри штучного інтелекту
Харківського національного
університету радіоелектроніки Л.Е.Чала

Підпис доц. Чалої Л.Е. засвідчую:
Учений секретар Харківського
національного університету радіоелектроніки І.В. Магдаліна